The Extreme Value Model for Dengue Fever Rates Prediction in Nakhon Sawan Province, Thailand

Yupawadee Samranrit and Chom Panta

Abstract—**This study applied the theory of statistical analysis, Extreme Value Theory (EVT), with dengue fever cases reported between 2010 and 2019 from the Department of Disease Control Bureau of Epidemiology, Ministry of Health, Nakhon Sawan region. From the results of this study, it showed that the highest monthly recurrence values of dengue fever for Nakhon Sawan province in every 5, 10, and 20 years were 43.1725, 54.6569, and 68.1125 per 100,000 populations, respectively.**

Index Terms—**Extreme value, Dengue Fever, Generalized Extreme Value Distribution, Return level**

I. Introduction

Dengue fever is caused by infection with the dengue virus, which is spread by a female mosquito. In many countries around the world, the disease is a public health issue. Over the last 30 years, the disease has spread and become endemic in over 100 countries. The disease is most common in tropical countries, and more than 40 percent of the world's population (2500 million people) suffer from dengue hemorrhagic fever. In Thailand, a report on dengue hemorrhagic fever from the Bureau of Communicable Diseases, led by insects on October 22, 2019, showed that there were 105,190 cumulative cases and the morbidity rate was 158.65 per 100,000 populations. In 2018, there was an increase in the number of dengue fever patients when compared to the same period of 2018, which were 55 percent (1.5 times) and 114 cases died and a mortality rate of 0.11%. For Nakhon Sawan Province, the Provincial Public Health Office reported that the total number of dengue fever cases in 2019 was 1,033, with a morbidity rate of 101.52 cases per 100,000 populations without dead patients. According to the reference data, it can be concluded that dengue fever tends to be higher in cases at both provincial and national levels.

There is currently no specific treatment for dengue fever. Treatment, therefore, is palliative related to symptoms. The treating physician must understand the nature of the disease and provide close care to the patients. Patients must be in good medical care throughout the critical period especially for 24-48 hours when leaking of plasma. Therefore, having effective methods to control dengue fever is extremely important. Prediction of the growth of the disease is another way to control disease progression. A result of the forecasting can provide surveillance and knowledge of the likelihood of disease in the future. It can lead to effective planning, preparation, prevention, and treatment of relevant agencies. In a review of relevant literature, researchers have proposed a variety of methods for predicting morbidity rates. For example, [1] studied the prognostic model of the dengue fever epidemic in eight Northern provinces of Thailand. They found that the suitable model for analyzing time series was ARIMA (1,0,0) (1,0,0) [12]. The highest number of cases was from June to July of every year. The forecast range for the number of cases was between 1.08 – 59.84. with a time series accuracy measure of 32.0%. The Bureau of Communicable Diseases, led by insects (2020) had predicted dengue fever in 2020 by using dengue cases monthly data for the past 10 years from January 2010 to September 2019 through two methods, Exponential Smoothing Method, and Box-Jenkins Method. The study showed that epidemic forecasting results with ARIMA (1,1,2) (1,1,2) in Thailand were rising consecutively from 2019 to 2020 with 146,361 cases. Several pieces of research in the previous had shown that if time-series analysis was utilized in dengue prognosis, discrepancies were often found to be inconsistent with preliminary agreements. For example, the higher rate of morbidity data might be construed as unusual information and was eliminated for analysis. However, the maximum number of cases data was important to obtain in the analysis. For example, if the morbidity rate data is high may be construed as unusual information. and were eliminated for analysis. However, when considering the information, the maximum number of cases data is critical to the analysis. Extreme Value Theory (EVT) was developed in the 1920s [2] and has been used to model maximum temperatures using data obtained from Penang [3]. To our knowledge, applications of EVT in public health are scarce. In the first applied this method to predict extreme influenza mortality in the US [4]. More recently, a study applied EVT to predict the Rate of Influenza Patients in the Northeast of Thailand [5].

Therefore, resources must be prepared to prevent and handle the most effective treatment. To predict the recurrence rate of dengue fever in 5, 10, and 20 years using the study of the highest monthly incidence in Nakhon Sawan Province, the researcher is interested in applying extreme value theory (EVT) to the data with a generalized extreme value distribution (GEVD).

Manuscript received Dec 17, 2021; revised Oct 11, 2022.

This research was funded by Research and Development Institute, Nakhon Sawan Rajabhat University, Thailand, the fiscal year 2020.

Yupawadee Samranrit is a Lecturer of Mathematics and Statistics program, Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Thailand. (e-mail: yupawadee.s@nsru.ac.th)

Chom Panta is an Assistant Professor of Mathematics and Statistics program, Faculty of Science and Technology, Nakhon Sawan Rajabhat University, Thailand. (Corresponding author to provide e-mail: chom.p@nsru.ac.th)

II. DATA

This research employed the secondary data of dengue morbidity rates in Nakhon Sawan province reported from the Department of Disease Control, Bureau of Epidemiology, Ministry of Health. The data was collected from 15 districts between 2010 and 2019, totaling 120 months. This data would be analyzed using the extreme value model with R.

Table 1: The basic statistics of monthly report data of dengue fever cases

	Minimum value	Maximum value	Skewness	Mean	p-value of KS.test
The number of cases of hemorrhagic fever in Nakhon Sawan Province	9	634	1.530787	144.33	P<0.05

From Table 1, using basic statistics to analyze the monthly number of dengue fever cases, the lowest number of dengue fever cases was 9, and the highest number of dengue fever cases was 634. The skewness value was 1.530787. The p-value of Kolmogorov-Smirnov tests was less than 0.05 and monthly dengue incidence data in Nakhon Sawan province did not have a normal distribution ($p<0.05$). Therefore, it is suitable for applying the extreme value theory.

III. THEORIES AND CONCEPTS OF EXTREME VALUES

A. Central limit theorem

$\bar{x}$ is the mean of a random sample $x_1, x_2, \ldots, x_n$ of size n and comes from any distribution in which the mean is μ and the variance is σ^2. The distribution of $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ will converge to the standard normal distribution $N(0,1)$ when the random sample size is large $n \to \infty$.

B. Extreme value theory

To study the theory of extreme values, it has to employ the maximum value of the random sample of interest, X_n. Which M_n is the maximum value [6]

$$M_n = \max(X_1, X_2, \ldots, X_n)$$

The probability and cumulative distribution function of a random sample can be obtained as follows:

$$\Pr\{M_n \le z\} = \Pr\{X_1 \le z, \ldots, X_n \le z\}$$

$$= \Pr\{X_1 \le z\} \cdot \ldots \cdot \Pr\{X_n \le z\} = \{F(z)\}^n$$

In fact, the function $F(z)$ is an unknown real value. In practice, fundamental techniques will be applied to estimate values from the collected data to approximate the probability density function of M_n in the above equation. Although, it is difficult to bring the approximate function closer to the actual function, the equivalent approach to the central limit theorem can be approximated by the true function. Considering the distribution of $\frac{M_n - b_n}{a_n}$, where a_n and b_n are constants. This leads to finding the distribution of $\left\{F\left(\frac{M_n - b_n}{a_n}\right)\right\}^n$. It is found that when $(n \to \infty)$ does not enter a distribution as in the central limit theorem at $\frac{\overline{X}_n - \mu_n}{\sigma_n} \to N(0,1)$ but $\left\{F\left(\frac{M_n - b_n}{a_n}\right)\right\}^n \to G(z)$ where $a_n > 0$ and $b_n > 0$ are constants. $G(z)$ does not depend on the size of n and leads to the external types theorem.

C. The Generalized Extreme Value Distribution

Let x_i, where $i = 1,2,\ldots,n$ plays as independent random variables. The same cumulative distribution function, $F(x;\theta)$, assigns the maximum value to a random variable, $x_{(n)} = Max(x_1, x_2, \ldots, x_n)$, which is applied in the form of the generalized extreme value distribution. The equation consists of 3 parameters: μ is of Location parameter, σ represents of Scale parameter, and ξ acts of Shape parameter [7].

The generalization extreme values were introduced in 1955 by Jenkinson. It was divided into three functions, which are Gumbel Distribution, Fréchet Distribution, and Weibull Distribution. Later in 1978, Galambos created the Cumulative Distribution Function (CDF) of the generalized extreme value distribution, which details are as follows:

The first theorem, if there are sequence constants $a_n > 0$ and b_n then

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \xrightarrow{D} G(z) \quad (1)$$

where $n \to \infty$ for the function G is a non-degenerate distribution, then G is a member of the family. The distribution of generalized extreme values is given by the equation:

$$F(x;\mu,\sigma,\xi) = \exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right\} \quad (2)$$

On the condition of $\left\{z : 1+\xi\left(\frac{x-\mu}{\sigma}\right) > 0\right\}$, when $-\infty < \mu < \infty, \sigma > 0$ and $-\infty < \xi < \infty$, has a Probability Function (pdf) as follows:

$$f(x)=\frac{1}{\sigma}\left\{1+\xi\left(\frac{x-\mu}{\sigma}\right)\right\}^{(-1/\xi)-1}\exp\left\{-\left(1+\xi\left(\frac{x-\mu}{\sigma}\right)\right)^{-1/\xi}\right\} \quad (3)$$

For $1+\xi\left(\frac{x-\mu}{\sigma}\right)>0$

From equations (2) and (3), where $\xi=0$, $\xi>0$ and $\xi<0$ the generalized extreme value distribution is called the Gumbel Distribution, the Fréchet Distribution, and the Weibull Distribution, respectively.

D. Return Level

In hydrology, the size of a catastrophic event used in the design of various water resources engineering projects is often referred as the return level (Z_p), which is the location of the data. p is the probability of an event and T is the recurrence year. The correlation with the probability of an event is $T=\frac{1}{p}$. [8]

E. The return level of the generalization extreme value distribution

From Equation (1), the return level equation is obtained as Equation (4).

$$\hat{Z}_r=\mu-\frac{\sigma}{\xi}\left\{1-\left[-\log\left(1-\frac{1}{T}\right)\right]^{-\xi}\right\} \quad (4)$$

The variables $\hat{\mu}$, $\hat{\sigma}$ and $\hat{\xi}$ are used to estimate the maximum likelihood of the degree of return level.

IV. The Suitable Parameter Model of the Highest Monthly Dengue Morbidity Rate under Generalized Extreme Values of Nakhon Sawan Province

The differences of each pattern were analyzed by using the Likelihood ratio test to select the model parameters of Nakhon Sawan Province as shown in Tables 2 to 5.

Table 2: Number of Parameters (P) and Negative Log-Likelihood (NLL) of the Monthly Highest Dengue Fever Rate Data Model in Nakhon Sawan Province.

Type	P	NLL
Type 1: μ, σ, ξ is a constant	3	65.8837
Type 2: $\mu(t)=\beta_0+\beta_1 t$ by σ and ξ is a constant	4	54.2846
Type 3: $\sigma(t)=\exp(\alpha_0+\alpha_1 t)$ by μ and ξ is a constant	4	54.3489
Type 4: $\mu(t)=\beta_0+\beta_1 t$, $\sigma(t)=\alpha_0+\alpha_1 t$ by ξ is a constant	5	52.3312
Type 5: $\mu(t)=\beta_0+\beta_1 t$, $\sigma(t)=\exp(\alpha_0+\alpha_1 t)$ by ξ is a constant	5	58.3295

Table 3: Analysis results of model differences by using Likelihood ratio test to select model parameters of Nakhon Sawan Province.

Compare Type	Likelihood ratio	chi-square	p-value	Chosen Type
1 with 2	2.3987	3.8410	0.0115*	2
2 with 3	20.9921	3.8410	1.6682e-05*	3
3 with 4	15.1956	3.8410	4.3581e-06*	4
4 wieh 5	8.3459	0.0000	1.0000	4
1 wieh 4	17.6084	3.8410	3.025e-10*	4

Note: It was statistically significant at the 0.05 level.

Table 4: Parameter Estimation Values of the Highest Monthly Dengue Fever Rate Data Model in Nakhon Sawan Province.

Type	Parameter Estimation
Type 1: μ, σ, ξ is a constant	$\hat{\mu}$ = 0.4897 $\hat{\sigma}$ = 0.4121 $\hat{\xi}$ = 1.0894
Type 2: $\mu(t)=\beta_0+\beta_1 t$ by σ and ξ is a constant	$\hat{\beta}_0$ = 0.8901 $\hat{\beta}_1$ = 0.4363 $\hat{\sigma}$ = 0.7492 $\hat{\xi}$ = 0.7821
Type 3: $\sigma(t)=\exp(\alpha_0+\alpha_1 t)$ by μ and ξ is a constant	$\hat{\alpha}_0$ = -0.0954 $\hat{\alpha}_1$ =0.8095 $\hat{\mu}$ = 0.8897 $\hat{\xi}$ = 0.5654
Type 4: $\mu(t)=\beta_0+\beta_1 t$, $\sigma(t)=\alpha_0+\alpha_1 t$ by ξ is a constant	$\hat{\beta}_0$ = 0.8879 $\hat{\beta}_1$ = 0.6992 $\hat{\alpha}_0$ = 0.6361 $\hat{\alpha}_1$ = 0.1172 $\hat{\xi}$ = 0.2607
Type 5: $\mu(t)=\beta_0+\beta_1 t$, $\sigma(t)=\exp(\alpha_0+\alpha_1 t)$ by ξ is a constant	$\hat{\beta}_0$ =1.0129 $\hat{\beta}_1$ =0.5542 $\hat{\alpha}_0$ = -0.1005 $\hat{\alpha}_1$ =1.0624 $\hat{\xi}$ = 0.2987

From Table 2, NLL of the data model of the highest monthly dengue incidence rate in Nakhon Sawan Province sorted in ascending order are as follows: Type 1 (-65.8837), Type 5 (-58.3295), Type 2 (-54.2846), Type 3 (-54.3489) and Type 4 (-52.3312).

From Table 3, the Type 4 was selected as the best model of the highest monthly dengue morbidity rate data in Nakhon Sawan Province.

From Table 4, when Type 4 was selected as the best model of the highest monthly dengue morbidity rate data in Nakhon Sawan Province, the equation can be illustrated as follows.

$$X \sim \mathrm{GEV}\left(\mu(t) = 0.8879 + 0.6992\mathrm{t}, \sigma(t) = 0.6361 + 0.1172\mathrm{t}, \hat{\xi} = 0.2607\right)$$

Table 5: Type and Parameter Estimation Values of the Highest Monthly Dengue Fever Rate Data Model in Nakhon Sawan Province: Divided by District.

District	Type and Parameter Estimation Values
Phayuha Khiri	Type 1: $\hat{\mu} = 0.8971$ $\hat{\sigma} = 0.5402$ $\hat{\xi} = 0.7657$
Chum Ta Bong	Type 1: $\hat{\mu} = 0.3214$ $\hat{\sigma} = 0.2527$ $\hat{\xi} = 0.6166$
Krok Phra	Type 2: $\hat{\beta}_0 = 0.2598$ $\hat{\beta}_1 = -0.0314$ $\hat{\sigma} = 0.1753$ $\hat{\xi} = 0.5859$
Mae Poen	Type 2: $\hat{\beta}_0 = 0.6189$ $\hat{\beta}_1 = 0.4991$ $\hat{\sigma} = 0.3812$ $\hat{\xi} = 0.5439$
Nong Bua	Type 2: $\hat{\beta}_0 = 0.7112$ $\hat{\beta}_1 = 0.4031$ $\hat{\sigma} = 0.6632$ $\hat{\xi} = 0.4109$
Phaisali	Type 3: $\hat{\alpha}_0 = 0.1912$ $\hat{\alpha}_1 = -0.0241$ $\hat{\mu} = 0.1918$ $\hat{\xi} = 0.4761$
Mae Wong	Type 3: $\hat{\alpha}_0 = 0.6923$ $\hat{\alpha}_1 = 0.1185$ $\hat{\mu} = 0.6549$ $\hat{\xi} = 0.7821$
Mueang Nakhon Sawan	Type 4: $\hat{\beta}_0 = 0.9819$ $\hat{\beta}_1 = 0.7611$ $\hat{\alpha}_0 = 0.5198$ $\hat{\alpha}_1 = 0.3109$ $\hat{\xi} = 0.2551$
Takhli	Type 4: $\hat{\beta}_0 = 0.7119$ $\hat{\beta}_1 = 0.8104$ $\hat{\alpha}_0 = 0.4011$ $\hat{\alpha}_1 = 0.2019$ $\hat{\xi} = 0.7323$
Lat Yao	Type 4: $\hat{\beta}_0 = 0.7763$ $\hat{\beta}_1 = 0.5419$ $\hat{\alpha}_0 = 0.4301$ $\hat{\alpha}_1 = 0.2909$ $\hat{\xi} = 0.1228$
Chum Saeng	Type 4: $\hat{\beta}_0 = 0.3619$ $\hat{\beta}_1 = -0.1125$ $\hat{\alpha}_0 = 0.1174$ $\hat{\alpha}_1 = 0.2319$ $\hat{\xi} = 0.1907$
Tak Fa	Type 4: $\hat{\beta}_0 = 0.8702$ $\hat{\beta}_1 = 0.6244$ $\hat{\alpha}_0 = 0.4808$ $\hat{\alpha}_1 = 0.2466$ $\hat{\xi} = 0.1913$
Banphot Phisai	Type 4: $\hat{\beta}_0 = 0.4324$ $\hat{\beta}_1 = 0.4495$ $\hat{\alpha}_0 = 0.5119$ $\hat{\alpha}_1 = 0.1783$ $\hat{\xi} = 0.3464$
Kao Liao	Type 5: $\hat{\beta}_0 = 0.5406$ $\hat{\beta}_1 = 0.6011$ $\hat{\alpha}_0 = 0.0938$ $\hat{\alpha}_1 = 0.5445$ $\hat{\xi} = 0.6817$
Tha Tako	Type 5: $\hat{\beta}_0 = 0.3921$ $\hat{\beta}_1 = 0.2406$ $\hat{\alpha}_0 = 0.2991$ $\hat{\alpha}_1 = 0.4624$ $\hat{\xi} = 0.6422$

From Table 5, the Type 4 was selected the most as the best model 6 district are Mueang Nakhon Sawan, Takhli, Lat Yao, Chum Saeng, Tak Fa and Banphot Phisai.

Considering the suitability of the 3 formats shown above can be represented as a graph in Figure 1 and Figure 2.

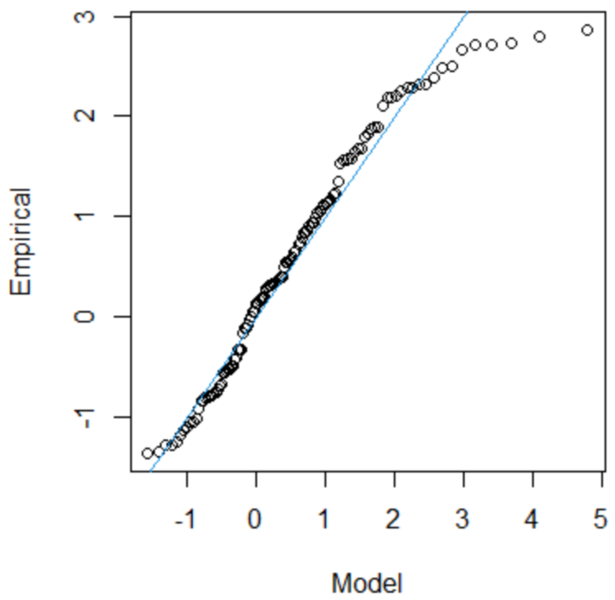

Figure 1: Quantile plot of generalized extreme values model with data on the highest monthly dengue incidence rate in Nakhon Sawan province.

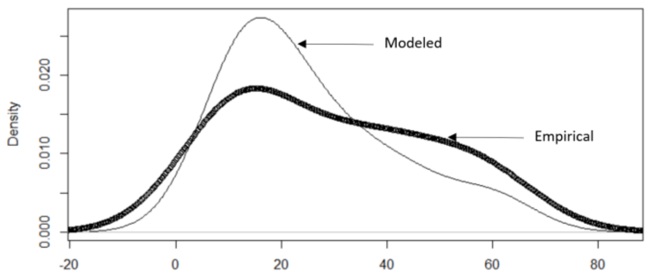

Figure 2: Density plot of generalized extreme values model with data on the highest monthly dengue incidence rate in Nakhon Sawan province.

From Figure 1 and Figure 2, the graph performing was in a straight line. The picture shows both data in the same distribution. A generalized extreme value model was suitable for this dataset.

The estimation of the return level in the 5-, 10-, and 20-years recurrence cycles of the highest monthly dengue morbidity rates in Nakhon Sawan Province is as shown in the following Table 6.

Table 6 Estimated return level (Zr) and standard error for 5, 10, and 20 years, respectively of the highest monthly dengue morbidity rate in Nakhon Sawan Province

	Return Level		
	Z_5	**Z_{10}**	**Z_{50}**
MLE	43.1725	54.6569	68.1105
(s.e.)	6.324	10.428	12.193

From Table 6, the researchers used 10 years of actual dengue fever incidence rate data between 2010 and 2019 to make a comparison of model data to match with real data of dengue incidence rates of Nakhon Sawan Province.

The results show that the highest morbidity rate corresponds to June 2011 and corresponds to index 18.

When taking recurrence estimates for a 5-year recurrence cycle, the highest monthly dengue morbidity rate was 43.1725 per 100,000 populations with a probability of 0.2.

For the 10-year recurrence cycle, the highest monthly dengue morbidity rate was 54.6569 per 100,000 populations with a probability of 0.1.

The last for a 20-year recurrence cycle, the highest monthly dengue morbidity rate is 68.1105 per 100,000 populations with a probability of 0.05.

V. Conclusion and Discussion of the Research Results

A generalized extreme value distribution was analyzed to determine the optimal parameter model of the monthly peak dengue morbidity data. The pattern difference analysis was done using the Likelihood ratio test. For selection of the parameter model, it was found that Type 4: $\mu(t) = \beta_0 + \beta_1 t$, with $\sigma(t) = \alpha_0 + \alpha_1 t$ as ξ constant, was the best model of the highest monthly dengue morbidity data. Therefore, the equation could be written to describe the meaning as follows:

$$X \sim \mathrm{GEV}\left(\mu(t) = 0.8879 + 0.6992\mathrm{t}, \sigma(t) = 0.6361 + 0.1172\mathrm{t}, \hat{\xi} = 0.2607\right).$$

When determining the return level in the recurrence cycle under GEVD, the highest monthly dengue morbidity rate was 43.1725 per 100,000 populations with a probability of 0.2 in the 5-year recurrence cycle. For the 10-year recurrence cycle, the highest monthly dengue morbidity rate was 54.6569 per 100,000 populations with a probability of 0.1. Lastly, for a 20-year recurrence cycle, the highest monthly dengue morbidity rate is 68.1105 per 100,000 populations with a probability of 0.05. Show that the applicability of EVT to epidemiologic data. A GEVD was fitted to block maximal and was used to calculate estimates of return levels and of risks of exceeding a defined threshold value over given time periods. In accordance with a study applied EVT to predict monthly incidence of avian influenza cases and applicability of EVT in Public Health [9] [10].

Therefore, the results of this analysis can be used as a guideline for resource planning to prevent disease outbreaks and cope with the most effective treatment. In addition, it might also be used to define target areas for planning and surveillance of dengue fever in Nakhon Sawan Province.

To conclude, EVT can be applied to epidemiological data, public health work, or data with high morbidity. It is an interesting alternative for predicting morbidity and mortality.

A limitation of this approach is the influence of climate uncertainties, such as temperature and humidity, that could affect the probabilistic value of predicted morbidity and mortality. For that reason, other suitable analytical methods and models are needed.

References

[1] N. Khantikul, P. Sudathip, A. Saejeng, R. Tipmontree, and W. Suwonkerd, "Social and Environmental Factors Affect Dengue Hemorrhagic Fever Epidemics in Upper Northern Thailand." Lanna Public Health Journal vol. 9, no.1, 2013, pp 23-36.

[2] Coles S. An introduction to statistical modeling of extreme values. Springer-Verlag; 2001.

[3] Husna B. Hasan, Noor Fadhilah B. Ahmad Radi, and Suraiya B. Kassim, "Modeling of Extreme Temperature Using Generalized Extreme Value (GEV) Distribution: A Case Study of Penang," Lecture Notes in Engineering and Computer Science: Proceedings of The World Congress on Engineering 2012, WCE2012, 4-6 July, 2012, London, U.K., pp181-186.

[4] Lee HC, Wackernagel H. Extreme values analyses of US P&I mortality data under consideration of demographic effects. Centre de géosciences. Ecole des Mines de Paris, Fontainebleau, France, 2007.

[5] W. Hencharoen, and P. Busababodin, "Extreme Value Modeling for the Rate of Influenza Patients in the Northeast of Thailand" Burapha Science Journal, vol. 22, no.3, 2017, pp 123-140.

[6] Coles, S. and Nadaraja, S., An Introduction to Statistical Modeling of Extreme Values. Great Britain: Springer Varlag London Limited, 2001.

[7] P. Busababodin, Extremes Value Analysis with R. Maha SaraKham : Maha Sarakham University, 2016.

[8] P. Senapeng, and P. Busababodin, "Modeling of Maximum Temperrature in Northeast Thailand" Burapha Science Journal, vol. 22, no.1, 2017, pp 92-107.

[9] Chen J, Lei X, Zhang L, Peng B. "Using extreme value theory approaches to forecast the probability of outbreak of highly pathogenic influenza in Zhejiang China". PloS ONE, 10(2), 2015.

[10] Thomas, M., Lemaitre, M., Wilson, ML., Viboud, C., Yordanov, Y., Wackernagel, H., & Carrat, F., "Applications of Extreme Value Theory in Public Health". PLoS ONE, 11(7), 2016.